DOCUMENT RESUME

ED 265 167 TM 850 502

AUTHOR Cronin, Linda L.; Capie, William
TITLE The Influence of Scoring Procedures on Assessment Decisions and Their Reliability.
PUB DATE Mar 85
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; *Behavior Rating Scales; *Classroom Observation Techniques; Data Collection; Elementary Secondary Education; *Generalizability Theory; Interrater Reliability; *Scoring; Teacher Behavior; Teacher Certification; *Teacher Evaluation; *Test Reliability
IDENTIFIERS Behaviorally Anchored Rating Scales; *Teacher Performance Assessment Instruments

ABSTRACT

The purpose of this study was to compare the scoring of Teacher Performance Assessment Instruments (TPAI) indicators using discrete descriptors when some are considered "essential" with the scoring of these same indicators, and when no descriptors are considered essential. The two questions addressed in this study were: (1) To what extent does the use of essential descriptors affect the overall "pass-fail rate" for each competency? and (2) To what extent does the use of essential descriptors affect the dependability of the certification decision? Data was used from twenty-six teachers who volunteered to prepare a lesson plan portfolio and who allowed observers to come into their classes. Results of observations were scored using two different methods: (1) using essential descriptors according to the criteria anticipated when the instruments are used in certification; and (2) treating descriptors equally, with no essential designation. Analyses were conducted on the transformed data obtained using the essential and non-essential scoring systems. Results showed that the essential descriptors did not detract from the reliability of the measures; in fact, it was enhanced by them. Although the study requires replication with a larger number of teachers and more realistic conditions, the results were viewed as supportive of the essential descriptor scoring method. (LMO)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED265167

# The Influence of Scoring Procedures on Assessment Decisions and their Reliability

Linda L. Cronin
William Capie
University of Georgia

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. Cronin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

April, 1985

Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois

TM 850 502



BEST COPY AVAILABLE

# THE INFLUENCE OF SCORING PROCEDURES ON ASSESSMENT DECISIONS AND THEIR RELIABILITY

Combining observation data to make evaluation decisions is a difficult and troublesome process when the performances to be evaluated are as complex as is teaching. Typically, there are a number of dimensions included in the assessment. And, each may be observed by a number of different individuals on a number of different occasions. Little of the work in personnel psychology speaks to the problem, since uses of observation data in that field are far different from the problems of licensure and certification. A performance profile may be created as part of a periodic evaluation of personnel with no need for cut-offs, per se. And, when personnel evaluations are included in a promotion process, they are only one criterion among many with no scoring formula. The need for a scoring formula arises when there is a large set of candidates, each of whom must be screened with regard to specified criteria.

Teacher licensure, as in the professional certification process in Georgia, or merit certification in Florida, are examples where standardized scoring procedures are necessary. The kinds of measurements to be made are certainly a key factor in the scoring. In Florida's Performance Measurement System, for example, observers tally behaviors in each occurrence. Large numbers of desirable behaviors are considered to constitute an effective performance. Consequently, the scoring system provides for summing behaviors and awarding "quality points" in proportion to the number of instances observed. These points are summed, in turn, to create a grand sum which is the teacher's score. The


BEST COPY AVAILABLE

score is derived without regard for the overall profile of performance or particular areas of strength or weakness.

Scoring in Georgia's certification program has been based on a different set of assumptions. First, and most important, teachers must demonstrate satisfactory performances in a number of subareas. As a result there is a limit to the degree to which one kind of behavior can compensate for another. Initially, the Teacher Performance Assessment Instruments (TPAI) were constructed of Behaviorally Anchored Rating Scales (BARS) which described five gradations of performance.

In this original system, the TPAI consisted of a set of broad teaching competencies which were defined by second-order descriptions of behaviors called indicators. The indicators were sentence length statements scored on a scale from 1-5. In order to determine or assign a score to each indicator, indicators were defined by third-order descriptions of behavior called descriptors. Each descriptor was assigned a scale point value. After observing an appropriate sample of teaching performance, observers selected the descriptor scale point best representing the teaching performance observed for each indicator.

As the TPAI use increased, some of the problems inherent with BARS emerged. Most notably, the end-points were fairly distinct but the mid-points tended to be somewhat less clear. Inasmuch as the basic scoring purpose of the TPAI indicator was to reduce the performance to a single satisfactory/not satisfactory decision, other methodologies were explored from the outset. The most successful was a set of discrete descriptors, as distinct from the hierarchical descriptors in BARS. A sample of an indicator and

its discrete descriptors is shown in Figure 1. Observers made a dichotomous decision about each descriptor and then the scoring rules for the indicator were used to translate these to a single decision about the adequacy of the indicator.

The discrete descriptors helped to avoid some of the ambiguity in the BARS, since each descriptor is a distinct statement. However, they also carried some limitations in scoring. Since the four descriptors were considered equal, any one of them could be unsatisfactory and the teacher's performance on the indicator could still be satisfactory. This tension between the desirability of independent, clearly-stated descriptors and the desirability of the weighting implied in the hierarchical BARS was not resolved in the current edition of the TPAI.

When the revised TPAI was planned, there was an effort to improve clarity and reduce ambiguity throughout. These efforts led to the decision to eliminate the hierarchical BARS and replace them with discrete descriptors. However, the scoring system was modified substantially when certain descriptors were dubbed "essential." An indicator could not be scored acceptable unless all of its essential descriptors were scored acceptable. A sample of a revised indicator is included in Figure 2. This revised scoring methodology represented a combination of the desirable attributes of the hierarchical and discrete descriptor formats.

There was some expectation that the use of essential descriptors would make the scoring more difficult for teachers and, at the same time, reduce reliability somewhat. However, the magnitude of these effects could not be anticipated with no

performance data. The field-test of a preliminary version of the revised TPAI provided an opportunity to investigate the psychometric properties of the TPAI when indicators were scored with essential descriptors.

## PURPOSE

The purpose of this study was to compare the scoring of TPAI indicators using discrete descriptors when some are considered "essential" with the scoring of these same indicators when no descriptors are considered essential. The two questions addressed in this study were: (1) To what extent does the use of essential descriptors affect the overall "pass-fail rate" for each competency? and (2) To what extent does the use of essential descriptors affect the dependability of the certification decision?

## PROCEDURES

Data from twenty-six teachers were used in the analyses. The teachers were volunteers who agreed to prepare a lesson plan portfolio and have observers come into their classes. Each teacher was observed by four observers: an administrator, a peer, and two representatives from a Regional Assessment Center, each of whom observed independently.

The observation data consisted of 104 sets (26 teachers x 4 observers) of pass/fail decisions for each of the 142 descriptors in the revised TPAI. The results of these observations were scored using two different methods. The first scoring system was based on the use of essential descriptors, according to the criteria anticipated when the instruments are used in certification.

In this system, if all of the descriptors keyed as essential for an indicator received passing scores, then credit could also be given for passing scores on descriptors not keyed essential. If all of the descriptors keyed as essential for an indicator did not receive passing scores, then performance of descriptors not keyed essential could not be used to compute the indicator score.

In order to compute raw indicator scores, the number of descriptors receiving passing scores was totalled. These raw scores ranged from 1 (no descriptors successfully demonstrated) to 5 (all four descriptors successfully demonstrated) for each indicator. Figure 3 contains a sample of raw indicator scores for a hypothetical competency.

Next, these raw scores were compared with the minimum standard score set for each of the indicators. If an indicator's raw score was equal to or greater than the minimum level, that indicator was assigned a transformed score of 1 (acceptable). If the raw score was less than the minimum level, the indicator was assigned a transformed score of Ø (unacceptable). The result was a matrix of thirty-five 1's and Ø's for each observation of each teacher. In Figure 4, the data shown in Figure 3 have been transformed to reflect an acceptable/unacceptable decision for each indicator.

In the second method of scoring, all descriptors were treated equally, i.e. none of them were designated essential. Raw indicator scores consisted of the total number of descriptors scored acceptably and ranged from 1 to 5 as in the essential scoring system. Transformed scores weere determined exactly as they were in the essential scoring system.

## ANALYSIS

Two types of analyses were conducted on the transformed data obtained using the essential and non-essential scoring systems.

### Pass-Rate

Each of the twenty-six teachers was scored on all levels of the TPAI using both the essential and non-essential scoring systems. The portion of indicators scored acceptably by the observers was the score. The mean "score" was computed for both of the scoring methods.

### Dependability

Generalizability theory was used to plan the analyses of the field test data. Four facets were identified as important sources of variation in the performance data obtained: teachers; observers; observer-types; and performance indicators. The four facet design with observers nested within observer-type is identical to a three facet fully crossed design with teachers, observer-types, and performance indicators as the sources of variation. As a consequence, the simpler three facet model was used in all analyses.

For each analysis, teachers were treated as facets of differentiation and observer-type and performance indicators within competencies were treated as facets of generalization. All facets were regarded as random in the analysis design.

A dependability coefficient ($\phi$) was calculated to assess the dependability of the data for making judgements about teacher performance relative to a standard ($\lambda$).



BEST COPY AVAILABLE

According to Brennan (1978):

$$\hat{\phi}(\lambda) = \frac{\hat{\sigma}^2(T) + (\bar{\chi} - \lambda)^2 - \hat{\sigma}^2_{\bar{\chi}}}{\hat{\sigma}^2(T) + (\bar{\chi} \quad \lambda)^2 - \hat{\sigma}^2_{\bar{\chi}} \hat{\sigma}^2_{\Delta}}$$

$\hat{\sigma}(T)$ = the variance attributable to the facet of differentiation:

$\hat{\sigma}^2_{\bar{\chi}}$ = the variance components of the observed mean scores;

$\bar{\chi}$ = the mean score;

$\lambda$ = the cutoff score;

$\hat{\sigma}^2_{\Delta}$ = variance component due to error

## RESULTS

The portion of indicators mastered by the twenty-six teachers across four observers is shown in Table 1. With no essential descriptors, the score was .81 or 81 percent. When some descriptors were designated essential, the score dropped to .77 or 77 percent.

Insert Table 1 about here

The variance components generated in the generalizability analysis are displayed in Table 2.

Insert Table 2 about here



BEST COPY AVAILABLE

Two coefficients were derived in each generalizability analysis-the generalizability coefficient ($\rho^2$) and the dependablility coefficient ($\phi$ ($\lambda$)). With no essential descriptors, these values were .65 and .89 respectively. When descriptors were designated essential, the two values were .68 and .91, respectively. These results are included in Table 3.

Insert Table 3 about here

## DISCUSSION

Consideration of these results must be made in light of the context in which the study was done. The TPAI which was used was a preliminary field-test edition which had had limited use prior to the study. Furthermore, the observers had had little or no training in its use and interpretation. The RAC members had participated in reviews of earlier drafts and had a brief orientation meeting. The school site observers may have had an orientation meeting. Such arrangements were tolerable in field-testing since the assessments were followed by extensive debriefing to identify problems which required attention in instrument revision and/or in training.

The difference in the mean performance levels using the two scoring systems was expected since the essential descriptors were, in essence, making the requirements more specific. More than likely, it will diminish as the instrument is used for certification and teachers attend to the criteria more systematically.

The coefficients computed were not analagous to those computed in field tests of the current TPAI. Those analyses, which yielded $\rho^2$ values in excess of .8, were different in three important ways. First, the raw scores were used in the analyses and they had a range of 1-5 whereas these unacceptable/acceptable scores can only be 1 or 0. Second, these earlier studies used a design that included an occasion facet and both it and the indicators were considered fixed. Finally, the earlier studies involved a more refined instrument, trained observers, and beginning teachers.

All of these factors would tend to increase teacher variance or decrease error variance and, therefore, increase the magnitude of the generalizability coefficient. In light of these factors the $\rho^2$ values near .7 were considered very encouraging.

To some extent the values of the coefficients may have been surpressed by a lack of homogeneity in the indicators. The relatively large variance component associated with the indicators supports this interpretation.

The two scoring procedures were not equally reliable, but the results were not in the direction that had been anticipated. Because different scores on a single essential descriptor could result in different scores when an indicator score was computed, it was anticipated that essential descriptors would diminish reliability. Undoubtedly there were disagreements between observers on decisions about essential descriptors. However, this type of error was offset by increased variance in teacher performance that was associated with the essential descriptor scoring system. This conclusion can be supported by comparing the

relative magnitude of the Teacher and Teacher by Observer by Indicator variance components in Table 2. This is a situation where objectivity and the ability to differentiate teachers can be seen to be quite different concerns.

The increased magnitude of the reliability coefficients may be an indication of the validity of the "essentialness" of the essential descriptors. To the extent that the coefficients measure internal consistency, the higher values associated with essential descriptors suggest a better measure of the construct of "teaching" when this scoring system is used. However, the validity will await a more appropriate study.

## CONCLUSIONS

The generalizability coefficient of .68 was somewhat lower than the analagous value associated with the current TPAI. However, the result is considered satisfactory in light of the tentative instrument used, the lack of training, and the less conservative analysis employed in previous studies.

The essential descriptors did not detract from the reliability of the measures; in fact it was enhanced by them. This surprising finding was viewed with considerable relief and a good bit of caution given the uncertain stability of variance components (Tobin and Capie, 1981). Although the study will require replication with a larger number of teachers and more realistic conditions, the results are viewed as supportive of the essential descriptor scoring method.



BEST COPY AVAILABLE

REFERENCES

Brennan, Robert L. (1978). Extensions of generalizability theory to domain referenced testing, American College Testing Program, Iowa City.

Tobin, K., & Capie, W. (1981). An empirical investigation of the stability of variance components and generalizability coefficients derived from teacher performance data. A paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Indicator 35: Manages disruptive behavior among learners.

Descriptors

a. Behavior of the entire class is monitored throughout the lesson.

b. Learners do not interfere with the work of others or interact inappropriately often or for an extended period.

c. Learners who interact inappropriately or otherwise interfere with the work of others are identified and dealt with quickly ***or*** no learners interfere with instruction.

d. Learners who interact inappropriately or otherwise interfere with the work of others are identified and dealt with appropriately (e.g., firmly, with suitable consequences for situation, effectively, etc.) ***or*** no learners interfere with instruction.

Figure 1. Sample indicator and discrete descriptors.

Indicator 27: Implements activities in a logical sequence.

Descriptors

a. Lesson is initiated with an interesting introduction.

#b. Necessary lesson components are addressed.

#c. Lesson components are sequenced to provide a logical development of lesson content.

d. Lesson is closed appropriately.

#Tentative recommendations for essential descriptors

Figure 2. Revised indicator and essential descriptors.



BEST COPY AVAILABLE

Table 1

Portion of Indicators Mastered
(I=35)

| Scoring System | Score |
|---|---|
| No Essential Descriptors | .81 |
| Essential Descriptors | .77 |

Table 2

Variance Components for Different Scoring Systems
(T=26, O=4, I=35)

| Source | Non-essential Scoring | Essential Scoring |
|---|---|---|
| Teacher(T) | .009 | .010 |
| Observer(O) | .001 | .001 |
| Indicator(I) | .021 | .028 |
| TO | .014 | .014 |
| TI | .025 | .023 |
| OI | .001 | .001 |
| TOI | .085 | .101 |

Table 3
Reliability Coefficients for Two Different Scoring Systems
(T=26, O=4, I=35)

| Scoring System | $\rho^2$ | $\phi$ ($\lambda$) |
|---|---|---|
| No-Essential Descriptors | .65 | .89 |
| Essential Descriptors | .68 | .91 |

Sample Competency: Teacher Miss Hypothetical

| Indicator | Recommended Minimum Level | Raw Score RAC | Raw Score Peer | Raw Score Adm |
|---|---|---|---|---|
| 1 | 3 | 5 | 5 | 4 |
| 2 | 4 | 2 | 2 | 4 |
| 3 | 3 | 4 | 5 | 5 |
| 4 | 4 | 2 | 3 | 2 |

Figure 3. Sample set of raw indicator scores.

Sample Competency: Teacher Miss Hypothetical

| Indicator | Transformed Score RAC | Transformed Score Peer | Transformed Score Adm |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | Ø | Ø | 1 |
| 3 | 1 | 1 | 1 |
| 4 | Ø | Ø | Ø |

Figure 4. Sample set of transformed indicator scores.